AFRL-HE-AZ-TR-2006-0015-Vol III

Air Force Research Laboratory

Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study: Real-Time and Blind Expert Subjective Assessments of Learning

Brian T. Schreiber

Lumir Research Institute
195 Bluff Avenue
Grayslake IL 60030

Sara Elizabeth Gehr

The Boeing Company
6030 South Kent Street
Mesa AZ 85212-6061

and

Winston Bennett, Jr.

Air Force Research Laboratory
Warfighter Readiness Research Division
6030 South Kent Street
Mesa AZ 85212-6061

Volume III

July 2006

Final Report for March 2002 to July 2005

Human Effectiveness Directorate
Warfighter Readiness Research Division
6030 South Kent Street
Mesa, AZ 85212-6061

NOTICES

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-AZ-TR-2006-0015-Vol III

This technical report has been reviewed and is approved for publication.

// Signed //

WINSTON BENNETT JR.
Project Scientist

// Signed //

HERBERT H. BELL
Technical Advisor

// Signed //

DANIEL R. WALKER, Colonel, USAF
Chief, Warfighter Readiness Research Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
07/31/2006

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
March 2002 to July 2005

4. TITLE AND SUBTITLE
Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study: Real-Time and Blind Expert Subjective Assessments of Learning

5a. CONTRACT NUMBER
F41624-97-D-5000

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
Schreiber, Brian T.[1]; Gehr, Sara Elizabeth[2]; & Bennett, Winston Jr.[3]

5d. PROJECT NUMBER
1123

5e. TASK NUMBER
AS

5f. WORK UNIT NUMBER
03

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

1. Lumir Research Institute
195 Bluff Ave
Grayslake, IL 60030

2. The Boeing Co.
6030 South Kent Street
Mesa AZ 85212-6061

3. Air Force Research Lab
Warftr Readiness Rsch Div
6030 South Kent Street
Mesa, AZ 85212-6061

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Division
6030 South Kent Street
Mesa AZ 85212-6061

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL; AFRL/HEA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-AZ-TR-2006-0015 – Vol III

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This is the third volume of a five-volume report, Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study.

14. ABSTRACT
The current work documented in this report is the subject matter expert (SME) rating data from a large study examining the within-simulator learning benefits of Distributed Mission Operations (DMO) training as described in AFRL-HE-AZ-TR-2006-0015 Vol I, Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study: Summary Report. That study examined 76 participating F-16 four-ship teams in week-long DMO training exercises and compared beginning-of-week to end-of-week performance on mirror-image air combat scenarios. As a major assessment component to the overall study, the current work reports data we collected from SME ratings of pilot performance. Two SME rating methods were employed—real-time ratings (i.e., done during mission execution) and SME ratings captured later using a scientific "blind" protocol. Comparing SME ratings of performance on the mirror-image scenarios revealed highly significant performance increases as a function of DMO training. SME real-time and blind ratings were both found to be significantly higher for end-of-week scenarios. The fact that the real-time *and* blind rater results corroborated one another provides strong support that performance greatly improved as a function of DMO training. Lastly, though the real-time ratings are not as scientific as the blind ratings, the fact that the blind ratings revealed very similar results to the real-time ratings greatly increases our confidence for continuing to use real-time ratings to assess performance. Compared to implementing an ongoing blind rating system, this justification to use real-time ratings creates a significant logistical, time, and financial savings for future research.

15. SUBJECT TERMS
Blind ratings; Distributed Mission Operations; DMO; MEC; Mission Essential Competencies; Networked environments; Performance assessment; Real-time ratings; Training effectiveness; Warfighter training

16. SECURITY CLASSIFICATION OF: Unclassified			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr. Winston Bennett Jr
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED	32	**19b. TELEPHONE NUMBER** *(include area code)*

This page intentionally left blank

Table of Contents

List of Figures and Tables

EXECUTIVE SUMMARY

In AFRL-HE-AZ-TR-2006-0015, Vol I, Distributed Mission Operations Within-Simulator Training Effectiveness Baseline Study: Summary Report, the authors reported numerous different data sources converging on the highly positive in-simulator training effectiveness of the Air Force Research Laboratory's Distributed Mission Operations (DMO) environment in Mesa, AZ. The current report documents only the subjective expert rating data from that study, but examines the data in more detail. We examined participating F-16 pilots in week-long DMO training exercises and compared beginning-of-week to end-of-week performance on mirror-image scenarios. To evaluate performance, we collected subject matter expert (SME) ratings of pilot performance via two methods. As is most common, we collected SME ratings in real-time (i.e., while SMEs were observing mission execution). Additionally, due to potential biases in the data (from SMEs knowledge of team experience and day of week), we collected additional SME ratings post-hoc according to a blind rater protocol. The blind rater protocol thus ensured scientific controls were in place and more straightforward conclusions could be drawn.

In conjunction with a computer-generated threat system and an instructor operator station, the DMO research environment in Mesa, AZ consisted of four high-fidelity F-16 simulators and one high-fidelity Airborne Warning and Control System simulator. From January 2002 to October 2004, participating F-16 teams flew over 40 total scenarios according to a five-day syllabus, book-ended on Monday and Friday by mirror-image point defense air combat benchmark scenarios. Comparing SME ratings of performance for 37 teams from the benchmarks revealed highly significant performance increases as a function of DMO training. SME real-time ratings for briefings/debriefings and for mission execution were found to be significantly higher on Friday than on Monday [$F(1, 28) = 97.22$, $p < 0.001$ and $F(1, 47) = 150.86$, $p < 0.001$, respectively]. The scientifically collected "blind" ratings of mission execution were also significantly higher for Friday missions, $F(1, 35) = 14.588$, $p = .001$.

The fact that the real-time and blind rater results corroborated one another provides strong evidence that the pilots did greatly improve skills as a function of DMO training. And, performance in the simulator was, to a very large statistical extent, significantly better on Friday across rated constructs. The effectiveness is further reinforced by our analysis of the other datasets from the overall study, (Schreiber, Stock, & Bennett, 2006b; Schreiber, Rowe, & Bennett, 2006d). Lastly, the fact that the blind ratings revealed very similar results to the real-time rating results greatly increases our confidence for continuing to use real-time ratings to assess performance. Compared to implementing an ongoing blind rating system, this justification to use real-time ratings creates a significant logistical, time, and financial savings for future research, but evidence from factor analyses suggests that the SMEs, using the current instrument, may not be sensitive enough to delineate performance among constructs.

ACKNOWLEDGEMENTS

The opinions expressed within this report are those of the authors and do not necessarily reflect the views of the sponsoring/employing organizations. This work was funded by the U.S. Air Force under contract #F41624-97-D-5000.

The authors wish to thank the following individuals for their significant contribution (in alphabetical order):

LtCol (Ret.) Michael “Frenchy” France,
Mary Johnson,
Antoinette Portrey,
Traci Smith,
LtCol Steven “Simple” Symons, and
Dr. William A. Stock.

This page intentionally left blank

DISTRIBUTED MISSION OPERATIONS WITHIN-SIMULATOR TRAINING EFFECTIVENESS BASELINE STUDY: REAL-TIME AND BLIND EXPERT SUBJECTIVE ASSESSMENTS OF LEARNING, VOLUME III

INTRODUCTION

Schreiber and Bennett (2006) reported a Distributed Mission Operations (DMO) training effectiveness study, examining the largest DMO within-simulator training effectiveness database known to exist. In that effort, the authors reported numerous different data sources converging on the highly positive training effectiveness of an air combat DMO environment. As such, the paper's focus was to report the overall results stemming from the central hypotheses of each dataset. This report documents only the subjective expert rating data from that study, but reports the methods and results in greater detail. Here we discuss the importance of subjective data and report expert observer ratings done both in real-time during mission execution and ratings done post-hoc according to a blind rater protocol. Because this study was designed as part of a very large overall in-simulator effectiveness evaluation of the Mesa DMO environment, we focus the literature review, results, and discussion to those areas most pertinent to subject matter expert (SME) rater methodology/usefulness in general and specifically as applied at the Mesa DMO site.

Scientists rely on participant opinions and SME raters perhaps more than any other assessment methodologies for assessing complex tasks. Given DMO's complexity, it comes as no surprise then that DMO assessment evaluations often rest primarily upon the use of domain experts rating the performance of the warfighter participants. Domain experts are uniquely qualified and capable of identifying subtle cues that differentiate experienced operators in a complex task and do so in real-time during mission execution. SMEs can identify early mistakes that can have a domino effect on mission performance or evaluate constructs that have been historically difficult to assess by other methods (e.g., "listens"). SMEs are also able to summarize overall performance and articulate subtle skill differences in ways that no other assessment methodology can capture. Furthermore, using SMEs as assessors is a straightforward logistical solution to a complex measurement task.

Given that SMEs can provide unique performance insight, can perform judgments in real-time, and can provide a simple assessment solution for very complex measurement tasks, it is readily apparent why SME performance ratings are a default measurement choice for assessing performance in DMO. A number of studies specifically at the Mesa DMO site have found significant training effects using SME ratings as a primary dependent metric (e.g., Bennett, Schreiber, & Andrews, 2002; Crane, Robbins, & Bennett, 2000; Krusmark, Schreiber, & Bennett, 2004). Most recently and most relevant to protocols in the current work, Krusmark et al. (2004) used the exact same SME rater gradesheet as was employed in the current work. The authors evaluated a nearly identical demographic group (F-16 four-ship teams, but from 2000-2001), the participants flew in a slightly earlier configuration of the Mesa DMO environment, and the pilots participated in five days of DMO training—all similar to the methods in the current work. A significant effect of training was found for the 32 F-16 teams (148 pilots), but

the reliance on *real-time* ratings generated a few potential alternative explanations for the results—specifically, bias-related caveats that we will attempt to address in the current work.

For benefit of the larger DMO within-simulator training effectiveness study (Schreiber & Bennett, 2006), the subjective SME rating approach was employed for many of the previously mentioned benefits, and as such, it was specifically employed to serve as a complementary assessment to objective data collected. Objective data collected via a computer has many advantages over using SMEs and subjective ratings (e.g., counts events precisely, measures timings exactly, applies rule sets easily, etc.). Furthermore, subjective assessments may not be sensitive enough to reveal significant changes when objective outcome measures do indeed show significant change (Pohlmann & Reed, 1978). However, objective techniques cannot provide the unique insights/assessments previously mentioned. As a result, it is simply not possible to objectively assess all facets of performance within a DMO environment. The subjective assessments would therefore theoretically corroborate the objective assessments *and* round out the evaluations for those skill areas too difficult to capture via a computer. Ultimately, the single most important contribution of the current work to the larger, overall effectiveness study is, "Do the SME observers view participants' demonstrated skill improving as a function of training time spent in a DMO environment?"

We desired to use SME ratings to assess the more complex, higher order skills not conducive for computer-based assessment. For the overall DMO effectiveness evaluation, the authors attempted to develop objective measures for air combat outcomes and map many of the measures to the air superiority Mission Essential Competency (MEC) skills (Colegrove & Alliger, 2002). Schreiber, Stock, and Bennett (2006) were able to easily develop objective metrics for air combat outcomes (eight measures) and 56 measures assessing various MEC air superiority skills. However, many of the MEC air superiority skills simply did not lend themselves well to objective assessment, some due to inherent psychological objective measurement challenges (e.g., "listens") and some due to current limitations presented by limited data available in computer network traffic. For these MEC skills, SME observer ratings would have to be used. Unfortunately, the need to complete a DMO within-simulator effectiveness study was great and the study began (Schreiber & Bennett, 2006) before completion of a specific MEC-based subjective assessment system could be developed and implemented (a MEC-based subjective tool is in development; MacMillan, Entin, Morley, & Bennett, in press). Fortunately, the legacy subjective real-time assessment system had been in use at the Air Force Research Laboratory Mesa site with minor variations since an early DMO study examining situation awareness in the early 1990s (Waag & Houck, 1994; Waag, Houck, Greschke, & Raspotnik, 1995).

Examining data from this legacy subjective rating system, our most recent SME real-time rating research (Krusmark, Schreiber, & Bennett, 2004) generated a few reservations and prompted us to pause before relying solely on real-time SME subjective rating data as part of the overall DMO within-simulator training effectiveness study. The most disconcerting issue identified with the real-time ratings included a potential rater bias. That is, with real-time ratings, SMEs know the day of the week, experience of team, and they *may* also have a vested interested in showing pilot improvement as a function of training in the DMO site (the same site that employs the SME raters). These factors alone could account for rating improvements over a week. Additionally, it was discovered that the SMEs did not accurately record simple statistics such as kills, and

insufficient systematic variance among individually rated skills was found. Though there was an observed improvement in ratings as a function of DMO training time, the authors concluded that the subjective data alone could not discount multiple possible explanations for what the observed DMO performance improvement could be attributed to. Therefore, in addition to providing SME observer real-time rating data for the overall DMO effectiveness report, a need existed to perform a blind rating study to address the rater bias concerns. Ideally, blind SME ratings would corroborate real-time SME ratings, thus lending more credibility/validity to real-time ratings than would ordinarily be warranted.

CURRENT WORK

For the current work, we put the Mesa legacy subjective assessment form (see Appendix A) into real-time rating practice and performed an additional blind rating study, seeking to fulfill the following specific objectives:

1. Collect SME observer real-time ratings for the overall DMO within-simulator effectiveness evaluation (Schreiber & Bennett, 2006) that would provide an additional, complementary, and converging DMO in-simulator training effectiveness data source.
2. Conduct a blind rater study to provide an additional, scientifically proper DMO subjective effectiveness evaluation for the overall DMO within-simulator study.
3. Map the individual constructs from the legacy subjective assessment system to the air superiority MEC skills, if possible.
4. Use the results of the blind study to confirm/deny the validity of using real-time SME ratings for future work.
5. Assess inter-rater reliability.

The single most important objective of the current work is to answer the question, “Do the SME observers view participants’ demonstrated skill improving as a function of training time spent in a DMO environment?”

METHODS

Participants

Pilots. From January 1, 2002 to October 22, 2004, 76 F-16 teams participated in the overall DMO within-simulator training effectiveness research study (Schreiber & Bennett, 2006). An estimated 20% of the entire USAF F-16 worldwide *population* -- 384 pilots -- participated in that study. To participate in the study, operational F-16 squadrons vied for posted vacant DMO training research weeks at the Mesa research site, readily volunteering for available training research opportunities. As such, participants in this study were not randomly sampled. Of those 76 teams used in the overall study, 37 teams produced usable data for examination in this report, which investigates SME ratings of performance. The vast majority of the teams not included were eliminated due to a new, proof-of-concept subjective rating system (MacMillan et. al., in press) being tested and evaluated (in lieu of data collected with legacy system analyzed here).

This still left us with a sufficient sample from the overall effectiveness study--148 pilot participants from 37 teams, 146 male and 2 female with a mean age of 32.8, 10.4 years of military service, and a mean number of hours in an F-16 of 905.7.

SME Raters. Twelve SME raters were called upon to provide real-time ratings and three SME raters were asked to provide blind ratings. All SMEs were retired USAF F-16 pilots (Major or Lieutenant Colonel).

DMO Training Facility

A portion of the following information is from General Method in Schreiber, B. T. and Bennett, W. Jr. (2006).

In conjunction with a computer-generated threat system and an instructor operator station (IOS), the DMO research environment in Mesa, AZ consisted of four high-fidelity F-16 simulators and one high-fidelity Airborne Warning and Control System (AWACS) simulator. The F-16s, AWACS, and threat entities interoperated according to Distributed Interactive Simulation (DIS) standards (IEEE Standard for Distributed Interactive Simulation - Application Protocols, 1995) version 4.02 or version 6.0.

The high-fidelity F-16 Block 30 simulators utilized 360 degree out-the-window visual displays with either SGI Onyx II Reality Monsters or PC Nova IIs running Aechelon runtime software. The visual system used high resolution photo-realistic databases of the Sonoran desert overlaid on terrain elevation data of the region. The hardware was very nearly identical to that found in the actual F-16, as was the software (Software Capabilities Upgrade version 4). Depending on the type of mission to be flown, F-16 weapon load-outs for missions consisted of differing combinations of the gun, the Air Intercept Missile (AIM-9), the Advanced Medium Range Air-to-Air Missile (AMRAAM), and/or the Mk-82 and Mk-84 general purpose bombs. A high-fidelity Solipsys version 6 AWACS sensor simulation was also used to provide a more realistic environment.

The Automated Threat Engagement System (ATES) generated all adversaries. A computerized, real-time threat generation system, ATES operates on standard DIS networks, providing air-to-air (AA), air-to-ground (AG), and surface-to-air (SA) threats. The ATES incorporates aerodynamic modeling, atmospheric models, radar models, infra-red models, and data parameter tables for thrust, drag, lift, etc. For the current work, threat air models were the MiG-29, MiG-27/23, and Su-27 loaded with the AA-8, AA-10a, and AA-10c air-to-air missiles. Ground threats included the SA-2, SA-6, and SA-8, and anti-aircraft artillery (AAA). Threat aircraft performed maneuvers and/or scripted flight paths while reacting to the F-16s' maneuvers and weapons.

The debrief facility included five 50-inch plasma screens -- one for a God's eye view and one dedicated for each of the four F-16s. Each of the F-16 plasma screens presented four avionic displays from the F-16. The time synchronized replay included all communications and could be paused, fast-forwarded or rewound according to the lead pilot's desired use of the allotted debrief time.

As a training research installation striving to continually integrate and evaluate new training technologies, the DMO site at Mesa undergoes occasional upgrades to its simulation systems. Therefore, the DMO simulation environment was not constant for all participants in this study. Some examples of upgrades/changes to the environment during the 33-month data collection period include (but are not limited to):

- upgrading the visual databases in cockpits #3 and #4 to use the same photo-specific database used in cockpits #1 and #2,
- upgrading to eight visual channels,
- upgrading the radios,
- installing SCU-5 Situation Awareness Data Link (SADL) software,
- installing new ALQ-213 radar warning/electronic counter measure panels and 5100 power PC boards,
- adding smoke trails to missile fly-outs, and
- upgrading the brief/debrief facility with Portable Flight Planning Software version 3.2 and a sixth 50-inch plasma debrief display for AWACS.

Under most circumstances changing the apparatus during the course of a scientific study threatens the study's conclusions. However, for the current work, we viewed these changes in the DMO environment as highly desirable. Further explained, as a system of integrated technologies, all DMO environments will change and be constantly upgraded at every field location. By doing similarly in our experimental environment we more closely replicate the actual systems to which we aim to generalize. Furthermore, we argue that significant learning effects must be found in light of the additional error variance associated with updates/changes to the environment, because the DMO environments will undoubtedly undergo change. If a training effect is not found under these changing conditions, justification for DMO training does not exist.

Training Research Syllabi/Training Research Week

Table 1 shows a general timeline for each participating team. Participants arrived early Monday morning for five days of DMO participation. Upon arrival, participants were first given an inbrief on the objectives and procedures of DMO and the simulators, a tour of the facilities, and then given a research administrative session where they completed a demographic form, were assigned anonymous barcode identification numbers, and finally took the first Pathfinder exercise-- an electronic knowledge structure assessment tool.

Table 1 Participant General Timeline.

Session#	1	2	3	4	5	6	7	8	9
Day/time	Mon AM	Mon PM	Tues AM	Tues PM	Wed AM	Wed PM	Thur AM	Thur PM	Fri AM
Activity	Mesa Inbrief Admin Pathfinder Pilot Brief Fly Fam Pilot Debrief	Pilot Brief Fly 3 Benches+ Pilot Debrief Feedback Survey	Pilot Brief Fly 4-8 engmnts Pilot Debrief	Pilot Brief Fly 4-8 engmnts Pilot Debrief	Pilot Brief Fly 4-8 engmnts Pilot Debrief	Pilot Brief Fly 4-8 engmnts Pilot Debrief	Pilot Brief Fly 4-8 engmnts Pilot Debrief	Pilot Brief Fly 4-8 engmnts Pilot Debrief	Pilot Brief Fly 3 Benches+ Pilot Debrief Feedback Survey Reaction Survey Pathfinder Outbrief

Pilots participated in one of four very similar syllabi, each syllabus consisting of nine 3.5 hour sessions, beginning with session one on Monday morning and ending with session nine on Friday morning. There were two sessions each day of the 5-day training week, save Friday's single session. Each session entailed a one-hour briefing, an hour of flying multiple engagements of the same mission genre, and an hour and a half debriefing. The syllabi scenarios could be either offensive or defensive, but were all four F-16s versus X number of threats. Scenarios were designed with trigger events and situations to specifically train MEC skills (Symons, France, Bell, & Bennett, 2005). These syllabi were developed with traditional methods using full mission rehearsal scenarios across a spectrum of probable air-to-air missions and threats while increasing the complexity of the missions as the training research week progressed.

After completing the administrative tasks early Monday morning, each syllabus began with a familiarization session (session one) late Monday morning to orient pilots to DMO simulator environment specifics, such as visual ID characteristics and any switchology differences due to F-16 block number or F-16 mission software. The pilots required surprisingly little familiarity training. The hour allotted turned out to be more than enough familiarity time, as the high fidelity simulator layout and underlying simulation models closely resembled the actual aircraft and pilots quickly became comfortable with DMO simulator operation. Since the pilots readily and easily adapted to the simulation environment during the familiarization period, performance increases observed throughout the course of the subsequent sessions should be the result of learning/honing their skills and not learning "sim-isms" or other DMO idiosyncrasies.

Session two on Monday afternoon began with benchmarks (i.e., a "pre-test") used to measure pre-training performance. The training week ended with the "post-test" training benchmark session nine on Friday morning. The benchmark sessions consisted of flying three point defense engagements (see Figure 1). All benchmark point defense scenarios pitted the four participant F-16s and their AWACS controller against eight threats (six hostiles and two strikers) at a distance

greater than 40 nautical miles. During all benchmark scenarios, AWACS informed the F-16s (at long range to the threats) that there were six entities and that all six were already identified as hostile, thereby allowing the F-16s to shoot beyond visual range at those six entities. Regarding the two strikers, the AWACS operator could not "see" below 10,000 feet--the altitude under which the enemy strikers flew during all benchmarks. Therefore, the onus fell upon the F-16s to find any entities below 10,000 feet with their onboard radars and visually identify them before employing ordnance.

All benchmarks were designed to be equally complex according to the absolute complexity scoring scheme outlined by Denning, Bennett, and Crane (2002). Seven-point defense benchmark scenarios were developed, and the complexity analysis revealed that all benchmarks were indeed equally complex. Pilots flew in the same flight/cockpit assignment on Monday and Friday. Unbeknown to the pilots, for the Friday benchmarks, pilots flew the mirror-image of the three benchmarks that were flown on Monday. Strict data collection rules governed all benchmarks in order to maintain a realistic combat environment—i.e., no freezing or reloading entities, fuel always on, no reincarnating entities, no inserting new entities, real-time kill removal for all entities, no intervention/assistance from IOS operators, etc. Benchmarks terminated under one the following conditions: All F-16s dead, all air adversaries dead, enemy strikers reached their target, or 13 minutes elapsed time. During the course of the study, the vast majority of benchmarks terminated under one of the first three rules.

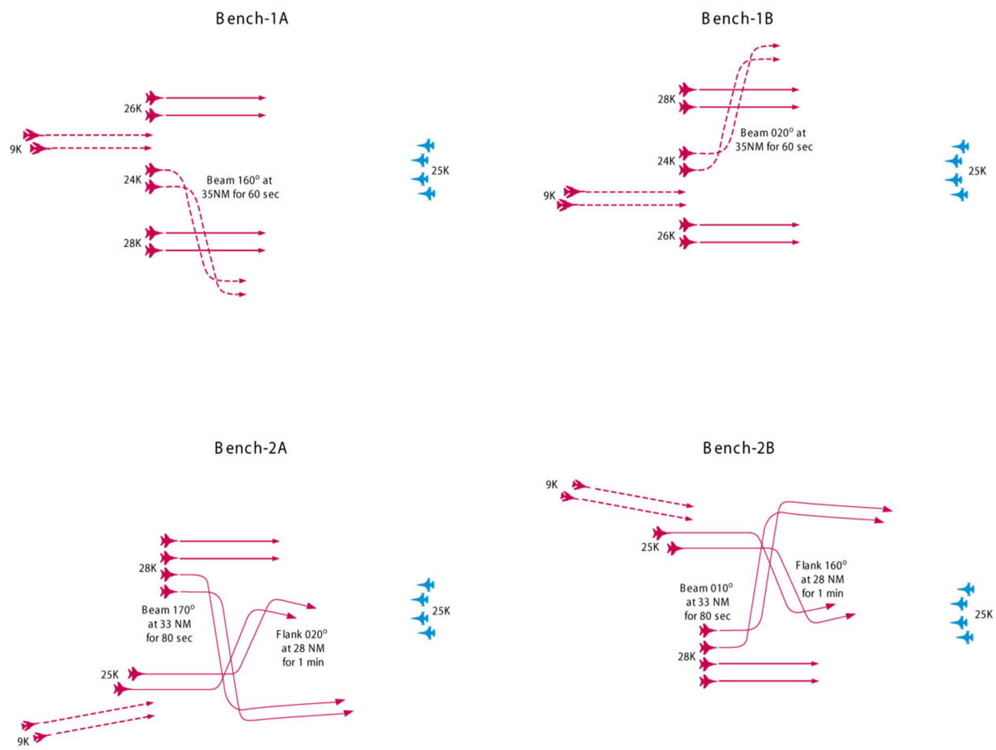

Figure 1 Example mirror-image point defense benchmark scenarios used for the pre- and post-test.

The participants' overriding goal for the point defense benchmark scenario was to prevent the enemy strikers/bombers from reaching the base – success being striker denial or kill. The second and third most important goals are to minimize friendly mortalities and maximize the adversary kills. The point defense benchmark scenarios were selected for examination in the present study

as pre- and post-test assessments because: (a point defense scenarios have very clear goals and measures of success, (b) all the benchmark engagements have equivalent levels of complexity, (c) three benchmark scenarios occur at the beginning and the end of the week-long DMO syllabus, (d) the same pilots in the same cockpit assignments perform the mirror-image (unknown to them) benchmark scenarios at the beginning and the end of the week, and (e) the benchmarks were flown under real-time kill removal and strict data collection rules.

The MEC-based building-block training began immediately after the benchmarks (with the remaining time during session two) on Monday afternoon and continued through the course of the week. Participating teams were exposed to four to eight full engagements per session. While these training sessions emphasized Defensive Counter Air scenarios (DCA), pilots also flew Offensive Counter Air (OCA) and air-to-ground missions. Usually, participating teams experienced about 35 training engagements between the Monday and Friday benchmarks, providing an intensive training curriculum. The building block training sessions progressed in complexity by increasing the number of threat aircraft, the type of threat aircraft, the threat aircraft reactivity/maneuver, and/or an increase in the vulnerability time.

Either after the last session on Thursday or on Friday morning, pilots took the second Pathfinder exercise and were given a DMO reaction rating form. After the last session on Monday and Friday, the team was also given a self-report feedback form with open-ended questions asking if they felt their objectives have been met and what facilitated or hindered their performance. Finally, before departure, teams were given a performance outbrief after their last set of benchmarks. This outbrief consisted of graphs for a number of the objective measures, revealing the team's performance.

Gradesheet

The legacy DMO gradesheet originates from air combat task analysis work done by Houck, Whitaker, and Kendall (1993). A gradesheet was derived and used in an early DMO study in 1993 to evaluate situation awareness (Waag, et al., 1995). Over the latter part of the 1990s, the gradesheet was slightly modified, eventually settling on its current form, which includes 57 broad indicators of 4-ship team performance (see Appendix A). Eight of these constructs evaluate the quality of the brief, 40 evaluate the pilots' performance during mission execution, and 9 evaluate the debrief. For reference, the grading criteria SMEs used in the current work to evaluate each construct were as follows:

N/A	Not applicable to this engagement
Grade D:	Dangerous
Grade O:	Performance indicates a lack of ability or knowledge
Grade 1:	Performance is safe, but indicates limited proficiency. Makes errors of omission or commission
Grade 2:	Performance is essentially correct. Recognizes and corrects errors
Grade 3:	Performance is correct, efficient, skillful, and without hesitation
Grade 4:	Performance reflects an unusually high degree of ability

Referring back to Table 1, for the real-time ratings the SME raters carried the gradesheet with them as they followed a particular team's training through the course of the week. Specifically,

each session to be evaluated by a SME (sessions 2-9) was evaluated using a single mission gradesheet, encompassing the brief, the multiple engagements, and the debrief for the entire session. Except for the rare computer failure, these gradesheets were electronic, filled out by the SMEs using a tablet-like computer (paper copies served as back-up). This gradesheet was used in real-time by the SMEs to grade the pilots' performance on the Monday and Friday benchmarks during sessions 2 and 9.

For the overall DMO within-simulator training effectiveness summary report (Schreiber & Bennett, 2006), we desired to assess not only the mission outcomes, but also as many MEC skills as possible. Using computer-based methods, the mission outcomes were comprehensively covered with objective techniques, but at this time it was possible to develop objective measures for only some of the MEC skills (Schreiber, Stock, & Bennett, 2006). Therefore, SME ratings for other skills would be used. A MEC-based subjective assessment system is in development (MacMillan, et al., in press), but it was not ready for formal use in the current work. Though the legacy assessment system employed here was developed before the MECs (and therefore was never intended to directly assess the MEC skills), we nonetheless anticipated that some of the gradesheet constructs might serve to assess some of the MEC skills. We provided two air combat SMEs, both experts in the MEC process, with a list of the gradesheet constructs and the MEC skills. Independently, they were asked to assign (if appropriate) a single MEC skill to a construct. As constructs or tasks usually have a many-to-many mapping with skills (i.e., skills tend to map to many different tasks/constructs in a domain), the SMEs left many mappings blank and only assigned a skill which clearly mapped to a construct. The result between SMEs was 13 identical skill mappings to gradesheet constructs. We chose not to pursue additional agreement procedures in attempt to cover more skills/constructs, as we were only interested in the clear, relatively unambiguous mappings to the legacy system (i.e., with a legacy system not originally designed to assess MECs, we felt only clear, non-debatable mappings should be used). The resultant 13 mappings are shown in Table 2 below.

Table 2 Mappings of MEC skills to legacy gradesheet mission execution constructs

Gradesheet Construct (mission execution only)	MEC air superiority skill
Radar Mechanics: El Strobe Control	Radar Mechanization
Radar Mechanics: Range Control	Radar Mechanization
Radar Mechanics: Azimuth Control	Radar Mechanization
Radar Mechanics: Utilizing Correct Mode	Radar Mechanization
Gameplan – Tactics	Selects Tactic
Tactical Intercepts: Formation	Maintains Formation
Tactical Intercepts: Targeting	Sorts Targets
Tactical Intercepts: Sorting	Sorts Targets
Tactical Intercepts: BVR launch and leave	Controls Intercept Geometry
Tactical Intercepts: BVR launch and react	Controls Intercept Geometry
Tactical Intercepts: Intercept Geometry	Controls Intercept Geometry
Post Merge Maneuvering	Executes Merge Gameplan
Communication: Radio Discipline	Prioritizes Communication

Blind Rater Protocol

For the blind rater evaluation, the benchmarks that were examined with the real-time ratings were put in random order. These randomized benchmarks (both Monday and Friday) were subjected to evaluation by SMEs according to a blind rater protocol. For evaluation, the SME raters had all the same pertinent resources to them that would have been available during real-time ratings—namely time-synchronized views of all the avionics for each pilot, a God's eye view, and all communication within the team. Dissimilar to the real-time ratings, the SMEs performing blind ratings also had the option to pause or rewind a replayed engagement (though they rarely did so). The raters watched the recorded scenarios in a random order, and—as was the sole purpose behind the blind rating effort in the current work--did not know the team (i.e., squadron), experience level, or if the benchmark that they were watching was from a Monday mission or a Friday mission. However, raters did know that each scenario was a benchmark from either session 2 or 9, and not a scenario from the training during the week (i.e., sessions 3-8). The SME raters rated only the performance of the pilots on the scenario or "flying" portions of the mission (i.e., 40 constructs on the gradesheet), and not on any of the brief or debrief categories, as the participant briefs/debriefs were not video recorded for later review. The raters watched and rated several benchmarks in a row, rating each one, and took breaks as needed.

RESULTS

Rater Reliability

From the blind rater effort, we had three raters who evaluated each scenario, allowing us to assess rater reliability. Because only one SME rated a given team for the real-time ratings, real-time rater reliability could not be assessed. We had data from all three blind raters for 51 benchmark scenarios. Each SMEs rating of all of the 40 areas graded on each benchmark was used for calculations. If there was missing data from one or more of the SMEs, that area of that benchmark was thrown out of the calculations. This resulted in 1752 points for which we had ratings from all three blind raters. Rater reliability was estimated using Ebel's method (1951,1967) of using mean squares (p.120). This resulted in 0.80 reliability of average ratings (i.e., reliability of the raters' average ratings) and 0.57 reliability of ratings (i.e., reliability of the raters at each construct on each benchmark).

Real-Time Ratings

A complete summary of real-time rating results is shown in Table 3. There were a total of 57 constructs to be rated, and the average ratings for a given construct ranged from .81-1.97 (mean = 1.36) on Monday to 2.10-2.93 (mean = 2.51) on Friday. For the real-time ratings, the ratings of the brief and debrief were analyzed separately from the engagement data because (a) we felt this was a natural assessment distinction from assessing actual simulator "flying," and (b) during a session, only one brief and one debrief period surrounded an hour of flying multiple SME-evaluated engagements (refer to Table 1). Therefore, there were less assessment data for the brief and debrief than for the engagements. For the brief and debrief data, there were 29 paired Monday and Friday benchmarks with complete data for all 17 brief and debrief constructs. These data showed that the SMEs rated participants significantly higher on Friday's brief and debrief (mean = 2.76) than on Mondays (mean = 1.75), $F(1, 28) = 97.22$, $p < 0.001$. For the

engagement portion of the gradesheet, there were 50 pairs of benchmarks for which we had complete data from a SME rater on all 40 “flying” constructs for both Monday and Friday. Over all engagement “flying” constructs, an analysis of variance showed that the gradesheet scores were significantly higher on Friday’s benchmarks (mean = 2.40) compared to Monday’s benchmarks (mean = 1.20), $F(1, 47) = 150.86$, $p < 0.001$. Follow-up t-tests revealed that for all 57 real-time rated constructs, Friday’s score was significantly higher than Monday’s ($p < 0.001$ for all).

Since all 57 constructs were significantly higher on Friday and the Krusmark, Schreiber, and Bennett (2004) study suggested a possible lack in measurement sensitivity, we performed an exploratory factor analysis on the real-time rating data. Separate principle component analyses were run on the Monday and Friday real-time ratings (results are shown in Appendix B). Scree plots revealed that just three factors seem to underlie both Monday and Friday ratings. A maximum likelihood procedure was run with a limit of three factors, again separately for Monday and Friday’s data. This showed that, of the three factors, there was some overlap in the first two factors, but the third factor was different on Monday and Friday. Additionally, a large number of the constructs had factor loadings above .35.

Blind Ratings

Complete blind rating results are shown in Table 3. There were 36 matched Monday and Friday benchmarks from the blind rater data. The blind rater data had some missing values where a SME had not rated a benchmark on one of the 40 constructs assessing engagement execution. If four or fewer data points (approximately 10%) were missing for a particular construct, the remaining data points were averaged, and that value put in place of the missing data. If five or more points were missing, that construct was not included in the analysis, and eight of 40 constructs did not meet this missing data criterion. For the remaining 32 constructs, average ratings for a given construct ranged from 1.33-2.86 (mean=1.85) for Monday to 1.92-2.91 (mean=2.33) for Friday.

These differences between the ratings on the Monday benchmarks and the Friday benchmarks were significant, $F(1, 35) = 14.588$, $p = .001$, even though the raters did not know what day’s benchmark they were watching. Follow-up t-tests revealed that 27 of the 32 constructs were significant ($p<0.05$). Two of the constructs (Radar Mechanics – range control and Radar Mechanics – Utilizing correct mode) approached significance ($p = 0.094$ and $p = 0.076$, respectively). Three of the constructs (E/F/N pole, ROE adherence, and ID adherence) were not significant ($p>0.1$).

Table 3 Real-time and blind rating results

	Monday real-time			Friday real-time			Monday blind			Friday blind		
Construct Rated	N	M	S.E.	N	M	S.E.	N	M	S.E.	N	M	S.E.
Brief: Mission Prep	29	1.97	0.13	29	2.90	0.11						
Brief: Developing Plan	29	1.69	0.13	29	2.83	0.13						
Brief: Organization	29	1.83	0.15	29	2.93	0.11						
Brief: Content	29	1.59	0.16	29	2.76	0.11						
Brief: Delivery	29	1.83	0.15	29	2.62	0.12						
Brief: Instructional Ability	29	1.62	0.15	29	2.55	0.11						
Brief: Sys Knowledge	29	1.86	0.15	29	2.66	0.10						
Brief: Overall Quality	29	1.55	0.13	29	2.90	0.08						
Radar Mech: El Strobe	48	1.15	0.09	48	2.35	0.13	36	2.06	0.14	36	2.42	0.12
Radar Mech: Range Control	48	1.56	0.09	48	2.54	0.11	36	2.08	0.13	36	2.39	0.11
Radar Mech: Azimuth Control	48	1.58	0.10	48	2.71	0.14	36	2.08	0.13	36	2.44	0.08
Radar Mech: Util.Correct Mode	48	1.44	0.10	48	2.56	0.11	36	2.08	0.15	36	2.42	0.10
Gameplan - Tactics	48	1.81	0.09	48	2.71	0.09	36	2.28	0.11	36	2.61	0.08
Gameplan: Execution	48	1.17	0.10	48	2.42	0.09	36	1.64	0.11	36	2.22	0.11
Gameplan: Adj..on-the-fly	48	0.94	0.10	48	2.31	0.13	36	1.33	0.11	36	2.06	0.12
TI: Formation	48	1.17	0.12	48	2.33	0.12	36	1.86	0.11	36	2.17	0.11
TI: Detection / Commit	48	1.69	0.09	48	2.88	0.11	36	2.25	0.10	36	2.69	0.08
TI: Targeting	48	1.56	0.13	48	2.58	0.12	36	2.22	0.14	36	2.61	0.10
TI: Sorting	48	1.21	0.12	48	2.44	0.12	36	1.83	0.12	36	2.34	0.13
TI: BVR launch and leave	48	1.08	0.11	48	2.21	0.12	36	1.92	0.13	36	2.46	0.11
TI: BVR launch and react	48	1.04	0.10	48	2.25	0.11						
TI: Intercept Geometry	48	1.33	0.10	48	2.21	0.10	36	1.56	0.12	36	1.97	0.07
TI: Low Altitude Intercepts	48	0.98	0.11	48	2.10	0.11						
Engagement Decision	48	1.13	0.12	48	2.23	0.11	36	1.81	0.12	36	2.25	0.11
Spike Awareness	48	1.25	0.10	48	2.52	0.13	36	1.86	0.18	36	2.41	0.11
E/F/N Pole	48	1.06	0.10	48	2.23	0.13	36	1.61	0.15	36	2.00	0.13
Egress / Separation	48	1.02	0.11	48	2.35	0.11	36	1.59	0.15	36	2.22	0.12
AAMD: RMD	48	0.98	0.11	48	2.40	0.12						
AAMD: IRCM	48	1.23	0.10	48	2.40	0.09						
AAMD: Chaff / Flares	48	1.17	0.11	48	2.60	0.09						
Contracts	48	1.13	0.10	48	2.29	0.11	36	1.75	0.11	36	2.31	0.11
ROE Adherence	48	1.27	0.12	48	2.31	0.12	36	2.80	0.20	36	2.91	0.22
ID Adherence	48	1.25	0.14	48	2.35	0.15	36	2.86	0.21	36	2.83	0.22
Post Merge Maneuvering	48	1.25	0.11	48	2.38	0.09	36	1.44	0.10	36	2.11	0.11
Mutual Support	48	0.90	0.10	48	2.23	0.14	36	1.50	0.13	36	2.11	0.12
Visual Lookout	48	1.08	0.08	48	2.21	0.11	36	1.36	0.13	36	2.25	0.12
Weapons Employment	48	1.27	0.11	48	2.50	0.10	36	2.33	0.11	36	2.69	0.10
Clear Avenue of Fire	48	1.73	0.12	48	2.56	0.12						
Comm: 3-1 Comm	48	0.98	0.06	48	2.25	0.10	36	1.64	0.11	36	1.92	0.12
Comm: Radio Discipline	48	1.06	0.08	48	2.29	0.10	36	1.75	0.10	36	2.17	0.12
Comm: GCI Interface	48	1.08	0.09	48	2.63	0.12	36	1.89	0.11	36	2.28	0.11
Fuel Management	48	1.65	0.12	48	2.60	0.13						

	Monday real-time			Friday real-time			Monday blind			Friday blind		
Flight Discipline	48	1.17	0.12	48	2.15	0.11	36	1.83	0.09	36	2.19	0.10
Situation Awareness	48	0.96	0.10	48	2.33	0.12	36	1.44	0.12	36	2.17	0.09
Judgment	48	1.02	0.11	48	2.25	0.12	36	1.50	0.11	36	2.17	0.09
Flight Leadership/Conduct	48	1.08	0.10	48	2.44	0.13	36	1.50	0.13	36	2.36	0.11
Briefed Objectives Fulfilled	48	0.85	0.09	48	2.44	0.12	36	1.47	0.10	36	2.31	0.12
Overall Engagement Grade	48	0.81	0.09	48	2.35	0.12						
Debrief: Organization	29	1.87	0.15	29	2.93	0.12						
Debrief: Reconstruction	29	1.97	0.17	29	2.86	0.12						
Debrief: Delivery	29	1.90	0.18	29	2.86	0.13						
Debrief: Analysis	29	1.84	0.17	29	2.93	0.11						
Debrief: Instr. Ability	29	1.71	0.17	29	2.72	0.11						
Debrief: ID Adherence	29	1.68	0.19	29	2.59	0.14						
Debrief: Flight Leadership	29	1.52	0.15	29	2.69	0.10						
Debrief: Miss Obj's Accomp	29	1.35	0.13	29	2.79	0.12						
Debrief: Overall Quality	29	1.68	0.16	29	2.86	0.10						

DISCUSSION

Results from the real-time ratings clearly reveal that SMEs view highly significant changes Monday to Friday as they observed performance over the course of the week. We anticipated this result, as the prior SME rating-based research performed at Mesa all revealed increases in SME real-time ratings over the course of a DMO training week (Bennett, et al., 2002; Crane, et al., 2000; Krusmark, et al., 2004). A concern and an unknown, however, was what would be the results of blind protocol ratings? Fortunately, the blind ratings, even though SMEs were not aware of time of week, showed very similar results to the real-time ratings. For the mission "flying" constructs that were rated (i.e., the constructs for which we had both real-time and blind ratings), the blind ratings revealed slightly tighter results between Monday and Friday than the real-time ratings (1.85-2.33 compared to 1.20-2.40). It is possible that the range and significance in real-time ratings are inflated somewhat over the blind ratings because of the potential real-time rating biases (or simply as a chance function of the smaller subset of blind scenarios rated). Additionally, ratings are, of course, relative, and do not reflect an absolute, quantifiable degree of improvement. Therefore, the differences are, for practical purposes, rather inconsequential and academic; both the real-time ratings and the blind ratings showed significant improvements as a function of DMO training. Addressing our primary objective for the current work, "Do the SME observers view participants' demonstrated skill improving as a function of training time spent in a DMO environment?" The data reported herein very clearly suggest, "Yes."

The fact that the real-time and blind rater results for the "flying" constructs corroborate one another and corroborate our objective results (Schreiber, Stock, & Bennett, 2006) provides us with several important conclusions. First, and most importantly, the two different datasets of SME ratings and the objective results reveal the same story—that the pilots performance greatly improved as a function of DMO training. And, performance in the simulator was, to a very large statistical extent, significantly better on Friday across a great many constructs. Additionally, though the real-time ratings are not as scientific as the blind ratings, the fact that the blind ratings revealed very similar results to the real-time ratings greatly increases our confidence for

continuing to use real-time ratings to assess performance. Compared to implementing an ongoing blind rating system, this justification to use real-time ratings creates a significant logistical, time, and financial savings for future research.

One concern does remain, however, with the ratings. Since the factor analyses relied on a relatively small sample, we consider the results to be exploratory. But, even so, with a large number of constructs providing fairly high factor loadings, the results may suggest that the instrument is not sensitive. Combined with the results from the Krusmark et al. (2004) study using the same gradesheet and also discovering a general lack in sensitivity, we postulate that the current SME rating instrument is likely only measuring a few true constructs and is not sensitive enough to delineate all the constructs listed on the gradesheet. Therefore, though there were many significant effects found across the constructs (both real-time and blind), interpreting any of these differences for the MEC skills-related constructs (or for the non-MEC skill constructs) is probably moot, as the larger issue across all the constructs may be a psychometric one with a few underlying factors. Additional research would be necessary to identify these constructs. Additionally, extensive evaluation of the new, MEC-based subjective assessment tool (MacMillan et al., in press) should also be undertaken.

Briefs/debriefs were only formally evaluated in the overall DMO within-simulator training effectiveness study by using the SME real-time ratings. Fortunately, based upon the demonstrated corroboration of the objective data and the blind ratings with the real-time ratings for the simulator "flying" performance, we can attribute more validity/credibility to the brief/debrief real-time ratings than we might have otherwise. Unsurprisingly, performance in this area was also significantly better on Fridays. Interestingly, AWACS controllers reported that skill acquisition was particularly strong during debriefs (Schreiber, Rowe, & Bennett, 2006). However, for all ratings in the current work (brief/flying/debrief), the significantly higher ratings on Friday, of course, reflect a *total effect of the entire training experience*. Thus, we do not know the degree to which the improvements from the training week are attributable to learning that took place during briefs, actual simulator time, or debriefs. Additional research is necessary to tease out the individual effects of each.

REFERENCES

Bennett, W. Jr., Schreiber, B. T., & Andrews, D. H. (2002). Developing competency-based methods for near-real-time air combat problem solving assessment. *Computers in Human Behavior,* 18(6), 773-782.

Colegrove, C. M. & Alliger, G. M. (2002, April). *Mission Essential Competencies: Defining combat mission readiness in a novel way*. Paper presented at the NATO RTO Studies, Analysis and Simulation (SAS) Panel Symposium. Brussels, Belgium.

Crane, P. M., Robbins, R., & Bennett, W. Jr. (2000). Using Distributed Mission Training to augment flight lead upgrade training. In *2000 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) Proceedings*. Orlando, FL: National Security Industrial Association (AFRL-HE-AZ-TR-2000-0111, ADA394919). Mesa, AZ: Warfighter Training Research Division.

Denning, T., Bennett, W. Jr., & Crane, P. M. (2002). Mission complexity scoring in Distributed Mission Training. In *2002 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) Proceedings*. Orlando, FL: National Security Industrial Association.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika,* 16 (4), 407-424. Reprinted in Mehrens, W. A. & Ebel, R. L. (1967) *Principles of educational and psychological measurement (*pp. 116-131). Chicago IL: Rand McNally and Co.,

Houck, M. R., Whitaker, L. A., & Kendall, R. R. (1993). *An information processing classification of beyond visual range air intercepts* (AL/HR-TR-1993-0061, AD A266 927). Williams Air Force Base, AZ: Armstrong Laboratory, Human Resources Directorate, Aircrew Training Research Division.

Institute of Electrical and Electronics Engineers. (1995). *IEEE Standard for Distributed Interactive Simulation - Application Protocols.* . NY: IEEE Publishing.

Krusmark, M., Schreiber, B. T., & Bennett, W. Jr. (2004). *The effectiveness of a traditional gradesheet for measuring air combat team performance in simulated Distributed Mission Operations*. (AFRL-HE-AZ-TR-2004-0090, AD A428 119). Mesa AZ: Air Force Research Laboratory, Warfighter Readiness Research Division.

MacMillan, J., Entin, E., Morley, R., & Bennett, W. Jr. (in press). Measuring team performance in complex and dynamic military environments: The SPOTLITE method. *The Journal of Military Psychology.*

Pohlmann, L. D., & Reed, J. C. (1978). *Air-to-air combat skills: Contribution of platform to initial training* (Rep. No. AFHRL-TR-78-53, AD A062 738). Williams Air Force Base, AZ: Air Force Human Resources Laboratory, Operations Training Research Division.

Schreiber, B. T. & Bennett, W. Jr. (2006). *Distributed Mission Operations within-simulator training effectiveness baseline study: Summary report.* (AFRL-HE-AZ-TR-2006-0015-Vol I). Mesa AZ: Air Force Research Laboratory: Warfighter Readiness Research Division.

Schreiber, B. T., Rowe, L. J., & Bennett, W. Jr. (2006). *Distributed Mission Operations within-simulator training effectiveness baseline study: Participant utility and effectiveness opinions and ratings.* (AFRL-HE-AZ-TR-2006-0015-Vol IV). Mesa AZ: Air Force Research Laboratory: Warfighter Readiness Research Division.

Schreiber, B. T., Stock W. A., & Bennett, W. Jr. (2006). *Distributed Mission Operations within-simulator training effectiveness baseline study: Metric development and objectively quantifying the degree of learning.* (AFRL-HE-AZ-TR-2006-0015-Vol II). Mesa AZ: Air Force Research Laboratory: Warfighter Readiness Research Division.

Symons, S., France, M., Bell, J., & Bennett, W. Jr. (2005). Linking knowledge and skills to Mission Essential Competency-based syllabus development for Distributed Mission Operations. (AFRL-HE-AZ-TR-1006-0041, AD A453 737). Mesa AZ: Air Force Research Laboratory, Warfighter Readiness Research Division.

Waag, W. L., & Houck, M. R. (1994). Tools for assessing situational awareness in an operational fighter environment. *Aviation, Space, and Environmental Medicine*, 65(Suppl. 5), A13-A19.

Waag, W. L., Houck, M. R., Greschke, D. A., & Raspotnik, W. B. (1995). Use of multiship simulation as a tool for measuring and training situation awareness. In AGARD Conference Proceedings 575 *Situation Awareness: Limitations and enhancement in the aviation environment* (AGARD-CP-575). (pp. 20-1 to 20-8). Neuilly-Sur-Seine, France: Advisory Group for Aerospace Research and Development.

ACRONYMS

AFRL	Air Force Research Laboratory
AA	Air-to-Air
AAA	Antiaircraft Artillery
AIM	Air Intercept Missile
AG	Air-to-Ground
AMRAAM	Advanced Medium Range Air-to-Air Missile
ATES	Automated Threat Engagement System
AWACS	Airborne Warning and Control System
DCA	Defensive Counter-Air
DIS	Distributed Interactive Simulation
DMO	Distributed Mission Operations
ID	Identification
IOS	Instructor Operator Station
MEC	Mission Essential Competencies
MK	Mark
OCA	Offensive Counter Air
SA	Surface-to-Air
SADL	Situation Awareness DataLink
SME	Subject Matter Expert
USAF	United States Air Force

This page intentionally left blank.

APPENDIX A: Brief/Mission/Debrief Gradesheets

Rater ID (last four): ___________________

Is there a benchmark engagement in this mission? **Yes** _____ **No** _____

DATE: __________ **DAY:** Mon Tue Wed Thurs Fri **TIME:** AM PM

Viper 1: ______________________ **Viper 2**: ________________________

Viper 3: ______________________ **Viper 4:** ________________________

General Mission Objectives (*check when briefed)*

____ No Morts ____ D/T/S all factor groups ____ Maintain mutual support

____ Effective use of GCI ____ 100% Valid shots ____ Good 3-1 Comm

Specific Mission Objectives (*list below*)

____

____

____

____

ROE *(please specify)* ____

____

____

Specific Gameplan/Tactics *(please specify)*

____

____

Overall Briefing Assessment	Grade
1. Mission Preparation	**NA D 0 1 2 3 4**
a. Developing Plan	**NA D 0 1 2 3 4**
2. Briefing	
a. Organization	**NA D 0 1 2 3 4**
b. Content	**NA D 0 1 2 3 4**
c. Delivery	**NA D 0 1 2 3 4**
d. Instructional Ability	**NA D 0 1 2 3 4**
3. Systems Knowledge	**NA D 0 1 2 3 4**
4. Overall quality of brief	**NA D 0 1 2 3 4**

Research Gradesheet

Team: _____Rater ID (last four): ______Pilot ID Number (5 Digit ID): ________

SCENARIO ID	Additional threats presented	Level of difficulty	Engagement Number
			1 2 3 4 5 6 7 8

Grading Criteria:

N/A: Not applicable to this engagement
Grade D: Dangerous
Grade O: Performance indicates a lack of ability or knowledge.
Grade 1: Performance is safe, but indicates limited proficiency. Makes errors of omission or commission
Grade 2: Performance is essentially correct. Recognizes and corrects errors.
Grade 3: Performance is correct, efficient, skillful, and without hesitation.
Grade 4: Performance reflects an unusually high degree of ability

Engagement Task	Grade	Notes
1. Radar Mechanics		
a. El Strobe Control	**NA D 0 1 2 3 4**	
b. Range Control	**NA D 0 1 2 3 4**	
c. Azimuth Control	**NA D 0 1 2 3 4**	
d. Utilizing correct mode	**NA D 0 1 2 3 4**	
2. Game plan / Tactics	**NA D 0 1 2 3 4**	
a. Execution	**NA D 0 1 2 3 4**	
b. Adjusting Plan On-The-Fly	**NA D 0 1 2 3 4**	
3. Tactical Intercepts		
a. Formation	**NA D 0 1 2 3 4**	
b. Detection / Commit	**NA D 0 1 2 3 4**	
c. Targeting	**NA D 0 1 2 3 4**	
d. Sorting	**NA D 0 1 2 3 4**	
e. BVR launch and leave	**NA D 0 1 2 3 4**	
f. BVR launch and react	**NA D 0 1 2 3 4**	
g. Intercept Geometry	**NA D 0 1 2 3 4**	
h. Low altitude intercepts	**NA D 0 1 2 3 4**	
4. Engagement Decision	**NA D 0 1 2 3 4**	
5. Spike Awareness	**NA D 0 1 2 3 4**	
6. E/F/N Pole	**NA D 0 1 2 3 4**	
7. Egress / Separation	**NA D 0 1 2 3 4**	

Grading Criteria:
N/A: Not applicable to this engagement
Grade D: Dangerous
Grade O: Performance indicates a lack of ability or knowledge.
Grade 1: Performance is safe, but indicates limited proficiency. Makes errors of omission or commission
Grade 2: Performance is essentially correct. Recognizes and corrects errors.
Grade 3: Performance is correct, efficient, skillful, and without hesitation.

Engagement Task	Grade	Notes
8. AAMD		
a. RMD	**NA D 0 1 2 3 4**	
b. IRCM	**NA D 0 1 2 3 4**	
c. Chaff / Flares	**NA D 0 1 2 3 4**	
9. Contracts	**NA D 0 1 2 3 4**	
10. ROE Adherence	**NA D 0 1 2 3 4**	
11. ID Adherence	**NA D 0 1 2 3 4**	
12. Post Merge Maneuvering	**NA D 0 1 2 3 4**	
13. Mutual Support	**NA D 0 1 2 3 4**	
14. Visual lookout	**NA D 0 1 2 3 4**	
15. Weapons Employment	**NA D 0 1 2 3 4**	
16. Clear Avenue of Fire	**NA D 0 1 2 3 4**	
17. Communication		
a. 3-1 Comm	**NA D 0 1 2 3 4**	
b. Radio Discipline	**NA D 0 1 2 3 4**	
c. GCI Interface	**NA D 0 1 2 3 4**	
18. Fuel Management	**NA D 0 1 2 3 4**	
19. Flight Discipline	**NA D 0 1 2 3 4**	
20. Situation Awareness	**NA D 0 1 2 3 4**	
21. Judgment	**NA D 0 1 2 3 4**	
22. Flight Leadership/ Conduct	**NA D 0 1 2 3 4**	
23. Briefed Objectives Fulfilled	**NA D 0 1 2 3 4**	

Objective Measures: Team Performance Statistics

Viper	AIM-120	AIM-9	Gun	Number of Kills	# Invalid	Explanation
1	**1 2 3 4 5 6**	**1 2 3 4**	**1 2**	**1 2 3 4 5 6 7 8**		
2	**1 2 3 4 5 6**	**1 2 3 4**	**1 2**	**1 2 3 4 5 6 7 8**		
3	**1 2 3 4 5 6**	**1 2 3 4**	**1 2**	**1 2 3 4 5 6 7 8**		
4	**1 2 3 4 5 6**	**1 2 3 4**	**1 2**	**1 2 3 4 5 6 7 8**		
TOTALS						

Survivors ___________ Morts ___________ Frats ___________

Overall Engagement Grade: 0 1 2 3 4

Grading Criteria:

N/A: Not applicable to this engagement
Grade D: Dangerous
Grade O: Performance indicates a lack of ability or knowledge
Grade 1: Performance is safe, but indicates limited proficiency.
Makes errors of omission of commission
Grade 2: Performance is essentially correct. Recognizes
and corrects errors
Grade 3: Performance is correct, efficient, skillful, and
without hesitation
Grade 4: Performance reflects an unusually high

Mission Evaluation Sheet

(please rate the debrief using the following criteria)

Debrief

Debriefing Task	Grade	Notes
1. Debriefing		
a. Organization	**NA D 0 1 2 3 4**	
b. Reconstruction	**NA D 0 1 2 3 4**	
c. Delivery	**NA D 0 1 2 3 4**	
d. Analysis	**NA D 0 1 2 3 4**	
e. Instructional Ability	**NA D 0 1 2 3 4**	
2. ID Adherence	**NA D 0 1 2 3 4**	
3. Flight leadership	**NA D 0 1 2 3 4**	
4. Mission Objectives Accomplished	**NA D 0 1 2 3 4**	
5. Overall quality of debrief	**NA D 0 1 2 3 4**	

APPENDIX B: Principle Component Analyses

(Note: Only those values with loadings >.35 are shown)

Gradesheet Construct	Factor 1 (Mon)	Factor 1 (Fri)	Factor 2 (Mon)	Factor 2 (Fri)	Factor 3 (Mon)	Factor 3 (Fri)
AAMD: Chaff / Flares	0.622			0.393	0.414	0.500
AAMD: IRCM	0.618			0.426	0.371	0.382
AAMD: RMD	0.766	0.352			0.387	0.642
Briefed Objectives Fulfilled		0.815			0.872	
Clear Avenue of Fire			0.460	0.650		
Communication: 3-1 Comm		0.380	0.461	0.463		0.488
Communication: GCI Interface		0.576	0.541	0.537	0.431	
Communication: Radio Discipline		0.407	0.410	0.434	0.356	0.572
Contracts	0.447	0.581	0.603			0.384
E/F/N Pole	0.546	0.423	0.555			0.740
Egress / Separation	0.654	0.491	0.522			0.746
Engagement Decision	0.802	0.606				
Flight Discipline		0.642	0.617		0.362	
Flight Leadership / Conduct	0.394	0.470	0.533	0.621		
Fuel Management	0.450		0.580	0.761		
Gameplan - Tactics	0.501			0.747		
Gameplan: Adjusting Plan On-the Fly	0.395	0.669	0.491		0.451	
Gameplan: Execution	0.520	0.668	0.361		0.465	0.378
ID Adherence		0.423	0.522			
Judgment		0.487	0.409	0.442	0.663	0.394
Mutual Support	0.561	0.668	0.381			
Overall Engagement Grade		0.865			0.869	
Post Merge Maneuvering	0.548	0.355	0.359	0.465		0.525
Radar Mechanics: Azimuth Control	0.510		0.482	0.858		
Radar Mechanics: Range Control	0.446			0.757		
Radar Mechanics: Utilizing Correct Mode	0.504		0.371	0.770		
ROE Adherence		0.564	0.676	0.482		0.357
Situation Awareness		0.612	0.441	0.545	0.639	
Spike Awareness	0.553			0.461		0.524
TI: BVR launch and leave	0.603	0.568	0.376			0.573
TI: BVR launch and react	0.672	0.514		0.428	0.467	0.471
TI: Detection / Commit	0.358			0.650		
TI: Formation	0.489	0.643	0.635			
TI: Intercept Geometry	0.472	0.574	0.541			
TI: Sorting	0.371	0.563	0.779	0.466		
TI: Targeting		0.370	0.633	0.778		
TI: Low Altitude Intercepts		0.582			0.510	
Visual Lookout	0.368	0.468	0.393		0.519	0.464
Weapons Employment			0.709	0.586		0.411

This page intentionally left blank.